第27卷 第3期 工 程 数 学 学 报 Vol. 27 No. 3
2010年06月 CHINESE JOURNAL OF ENGINEERING MATHEMATICS June 2010

文章编号:1005-3085(2010)03-0463-16

缺失数据下两线性模型中响应变量分位数差异的经验似然置信区间*

王历容[1]， 秦永松[2]， 白云霞[3]， 黎 玲[4]

(1- 娄底技师学院，娄底 417000； 2- 广西师范大学数学科学学院，桂林 541004;
3- 包头医学院药学系，包头 014040； 4- 广西师范大学漓江学院，桂林 541004)

摘 要: 总体差异的比较是医学，经济和教育领域中经常遇到的课题，本文讨论缺失数据情形下两线性模型中响应变量分位数差异的经验似然置信区间的构造。我们采用分数线性回归填补法对响应变量的缺失值进行补足，得到两线性回归模型的“完全”样本数据，在此基础上构造响应变量分位数差异的对数经验似然比统计量，在一定条件下证明了统计量的极限分布为加权 χ_1^2，并利用此结果构造分位数差异的经验似然置信区间。模拟结果表明在分数填补下得到的置信区间具有较高的覆盖精度。

关键词: 线性模型；分位数；分数线性回归填补；经验似然；置信区间

分类号: **AMS**(2000) 62G05; 62E20 **中图分类号**: O212.7 **文献标识码**: A

1 引言

经验似然是 Owen[1] 在完全样本下提出的一种重要的非参数统计推断方法，经验似然具有很多其他经典方法不具备的优点[2]，因此引起了众多统计学家的高度关注，随之被应用到完全样本下的多种模型和领域，并取得了不少研究成果，如文献 [3-5]。然而，在一些实际问题 (如市场调查，医学研究，民意调查等) 中，由于一些主观或客观方面的原因往往会造成样本数据缺失的现象，此时通常的统计方法不能直接应用，需要先对缺失数据进行处理，目前处理缺失数据最常用的方法是填补法，即对每个样本缺失值进行补足，构造“完全”样本，再利用通常的统计方法进行统计推断。如何将经验似然方法应用到不完全样本是一个重要的研究课题，近年来不少统计学者在这方面做了许多有益的探索[6-10]，其中文献 [10] 讨论了缺失数据下线性模型中响应变量分位数的经验似然置信区间，主要思想是采用分数填补法对缺失数据进行填补，得到总体的“完全”数据，再利用填补后的“完全”数据构造出线性回归模型中响应变量分位数的经验似然置信区间。响应变量缺失是实际中比较常见的一种情形，如在医学诊断和医学实验中，需要在固定时间点测量某个指标 (如体温)，但由于某种人为或客观原因，可能导致在某些时间点上某些测量指标缺失。分数填补法是一种多重填补法，最先由 Kim 和 Fuller[11] 提出，该填补方法可以减少随机填补带来的方差，所得区间精度比单一填补法更高[10]，因此该填补方法是一种较好的缺失数据填补方法。本文采用分数填补法将文献 [10] 中的部分结果推广到两个样本数据不完全的线性回归模型的情形，针对响应变量分位数差异的经验似然置信区间展开讨论，该问题的相关研究在目前的文献中还未被涉及，而且本文中所讨论的问题和模型具有广

收稿日期: 2007-12-14. 作者简介: 王历容 (1979年2月生)，女，硕士. 研究方向：数理统计.
*基金项目: 国家自然科学基金 (10971038)；广西科学基金 (0728092)；教育部留学回国人员科研启动资金 ([2004]527)；广西研究生教育创新计划资助项目 (桂学位 [2006]40).

泛的实际应用背景，例如，医学领域中，考察两种新药的疗效的差异是否明显，经济研究中，考察两个不同地区的居民生活水平的差异是否明显等，这些都是需要解决的实际课题。

对任意的分布函数$H(\cdot)$，定义其$q\,(0 < q < 1)$分位数$H^{-1}(q) = \inf\{x\,|\,H(x) \geq q\}$。用$F(\cdot)$和$G(\cdot)$分别表示总体$x$和$y$的分布函数，并用$\theta$记$x$的$q$分位数，$\Delta$记$x$和$y$的$q$分位数差异，即$\Delta = G^{-1}(q) - F^{-1}(q)$且$G^{-1}(q) = \theta + \Delta$。设$x$和$y$的样本分别满足以下两个相互独立的线性模型

$$x = u'\beta + A(u)\varepsilon, \quad y = v'\rho + B(v)\tau,$$

其中回归系数$\beta \in \mathbf{R}^p$，$\rho \in \mathbf{R}^q$，$A(u)$和$B(v)$为已知非负函数，随机误差变量ε和协变量u相互独立，τ和协变量v相互独立，且$E\varepsilon = E\tau = 0$，$0 < \mathrm{Var}\varepsilon = \sigma_\varepsilon^2 < \infty$，$0 < \mathrm{Var}\tau = \sigma_\tau^2 < \infty$。由于数据的缺失，我们得到以下不完全简单随机样本

$$Z_{x_i} = (x_i, u_i, \delta_{x_i}), \quad i = 1, \cdots, m; \qquad Z_{y_j} = (y_j, v_j, \delta_{y_j}), \quad j = 1, \cdots, n,$$

其中$\{u_i, i = 1, \cdots, m\}$和$\{v_j, j = 1, \cdots, n\}$可全部观测到，$\{x_i, i = 1, \cdots, m\}$和$\{y_j, j = 1, \cdots, n\}$有缺失，$x_i$只有在$\delta_{x_i} = 1$时才能观测到，而当$\delta_{x_i} = 0$时观测不到，$y_j$只有在$\delta_{y_j} = 1$时才能观测到，而当$\delta_{y_j} = 0$时观测不到，即

$$\delta_{x_i} = \begin{cases} 1, & \text{若 } x_i \text{ 不缺失}, \\ 0, & \text{若 } x_i \text{ 缺失}. \end{cases} \qquad \delta_{y_j} = \begin{cases} 1, & \text{若 } y_j \text{ 不缺失}, \\ 0, & \text{若 } y_j \text{ 缺失}. \end{cases}$$

用$Z_x = (x, u, \delta_x)$和$Z_y = (y, v, \delta_y)$分别表示$\{Z_{x_i}, i = 1, \cdots, m\}$和$\{Z_{y_j}, j = 1, \cdots, n\}$对应的总体。本文假定$\{x_i, i = 1, \cdots, m\}$和$\{y_j, j = 1, \cdots, n\}$均满足随机缺失机制(MAR)，即在给定$u$的条件下$\delta_x$与$x$相互独立，在给定$v$的条件下$\delta_y$与$y$相互独立，亦即

$$P(\delta_x = 1\,|\,x, u) = P(\delta_x = 1\,|\,u) = P_1(u), \quad P(\delta_y = 1\,|\,y, v) = P(\delta_y = 1\,|\,v) = P_2(v).$$

记

$$r_x = \sum_{i=1}^m \delta_{x_i}, \quad m_x = m - r_x, \quad r_y = \sum_{j=1}^n \delta_{y_j}, \quad m_y = n - r_y,$$

$$s_{r_x} = \{i : \delta_{x_i} = 1,\ i = 1, \cdots, m\}, \quad s_{m_x} = \{i : \delta_{x_i} = 0,\ i = 1, \cdots, m\},$$

$$s_{r_y} = \{j : \delta_{y_j} = 1,\ j = 1, \cdots, n\}, \quad s_{m_y} = \{j : \delta_{y_j} = 0,\ j = 1, \cdots, n\}.$$

类似于文献[10]，文中取窗宽$a = a_m > 0$，$b = b_n > 0$，以及核函数$K_1(\cdot)$和$K_2(\cdot)$，其中当$m \to \infty$时，$a \to 0$，当$n \to \infty$时，$b \to 0$。定义

$$G_1(t) = \int_{-\infty}^{t/a} K_1(u)\mathrm{d}u, \quad G_2(t) = \int_{-\infty}^{t/b} K_2(u)\mathrm{d}u,$$

$$\omega_1(x, \theta, \Delta) = G_1(\theta - x) - q, \quad \omega_2(y, \theta, \Delta) = G_2(\theta + \Delta - y) - q,$$

以上$\omega_1(x, \theta, \Delta)$和$\omega_2(y, \theta, \Delta)$即为文献[12]在完全样本情形下所定义的得分函数。为了采用分数线性回归填补法[10]对样本缺失值进行填补，我们先利用观测到的样本分别得到β和ρ的最小二乘估计

$$\hat{\beta}_r = \Big[\sum_{i=1}^m \Big(\frac{\delta_{x_i} u_i u_i'}{A^2(u_i)}\Big)\Big]^{-1} \sum_{i=1}^m \Big(\frac{\delta_{x_i} u_i x_i}{A^2(u_i)}\Big), \quad \hat{\rho}_r = \Big[\sum_{j=1}^n \Big(\frac{\delta_{y_j} v_j v_j'}{B^2(v_j)}\Big)\Big]^{-1} \sum_{j=1}^n \Big(\frac{\delta_{y_j} v_j y_j}{B^2(v_j)}\Big).$$

对于某个缺失样本 $x_i\,(i\in s_{m_x})$，从数据集 $\{A^{-1}(u)(x_j-u_j'\hat{\beta}_r),\, j\in s_{r_x}\}$ 中有放回地随机抽取 $k\,(k\in N,\, k$ 固定) 个数据，并依次记为 $\varepsilon_{il}^*\ (l=1,\cdots,k)$，再用 $u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*\ (l=1,\cdots,k)$ 作为 x_i 的填补值。对每一个缺失值均重复以上填补过程，从而得到填补后的 x 的“完全”样本

$$x_{I,i}=\big\{x_i,\, i\in s_{r_x};\ u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\ l=1,\cdots,k,\, i\in s_{m_x}\big\}.$$

对于得分函数 $\omega_1(x_i,\theta,\Delta)$, $i\in s_{m_x}$，则用

$$\frac{1}{k}\sum_{l=1}^{k}\omega_1\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\,\theta,\,\Delta\big)$$

来填补，同理可得 y 的“完全”样本

$$y_{I,j}=\big\{y_j,\, j\in s_{r_y};\ v_j'\hat{\rho}_r+B(v_j)\tau_{jl}^*,\ l=1,\cdots,k,\, j\in s_{m_y}\big\},$$

且用

$$\frac{1}{k}\sum_{l=1}^{k}\omega_2\big(v_j'\hat{\rho}_r+B(v_j)\tau_{jl}^*,\,\theta,\,\Delta\big)$$

来填补得分函数 $\omega_2(y_j,\theta,\Delta)$, $j\in s_{m_y}$，其中 $\tau_{jl}^*\in\{B^{-1}(v_i)(y_i-v_i'\hat{\rho}_r),\, i\in s_{r_y}\}$。

以上填补方法即为分数线性回归填补法，特别地，当 $k=1$ 时，即为通常的随机填补法。下面将利用填补后得到的“完全”样本 $\{x_{I,i},\ y_{I,j}\}$ 构造 Δ 的经验似然置信区间。我们将在第 2 节给出本文主要结果，第 3 节给出相关引理和主要结果的证明，第 4 节给出数值模拟结果。

2　主要结果

由于 x 和 y 的样本数据不完全，我们需要利用到以下新的得分函数

$$\omega_1(x_{I,i},k,\theta,\Delta)=\delta_{x_i}\omega_1(x_i,\theta,\Delta)+\frac{1-\delta_{x_i}}{k}\sum_{l=1}^{k}\omega_1\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta,\Delta\big),\quad i=1,\cdots,m,$$

$$\omega_2(y_{I,j},k,\theta,\Delta)=\delta_{y_j}\omega_2(y_j,\theta,\Delta)+\frac{1-\delta_{y_j}}{k}\sum_{l=1}^{k}\omega_2\big(v_j'\hat{\rho}_r+B(v_j)\tau_{jl}^*,\theta,\Delta\big),\quad j=1,\cdots,n,$$

其中

$$\frac{1}{k}\sum_{l=1}^{k}\omega_1\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta,\Delta\big)\quad 和\quad \frac{1}{k}\sum_{l=1}^{k}\omega_2\big(v_j'\hat{\rho}_r+B(v_j)\tau_{jl}^*,\theta,\Delta\big)$$

即为上节提到的缺失数据下相应的得分函数的填补值。类似于文献 [12]，定义经验似然函数

$$\prod_{i=1}^{m}p_i\prod_{j=1}^{n}q_j,$$

从而得到对数经验似然比统计量

$$R(\Delta)=\sup_{p_i>0,i=1,\cdots,m,q_j>0,j=1,\cdots,n,\theta}\Big\{\sum_{i=1}^{m}\log(mp_i)+\sum_{j=1}^{n}\log(nq_j)\Big\}=\sup_{\theta}R(\Delta,\theta),$$

其中

$$R(\Delta,\theta)=\sup_{p_i>0,i=1,\cdots,m,q_j>0,j=1,\cdots,n}\Big\{\sum_{i=1}^{m}\log(mp_i)+\sum_{j=1}^{n}\log(nq_j)\Big\},$$

并且 p_i, q_j 满足

$$\sum_i p_i=1,\quad \sum_{i=1}^{m}p_i\omega_1(x_{I,i},k,\theta,\Delta)=0,\quad p_i>0,\quad i=1,\cdots,m,$$

$$\sum_j q_j=1,\quad \sum_{j=1}^{n}q_j\omega_2(y_{I,j},k,\theta,\Delta)=0,\quad q_j>0,\quad j=1,\cdots,n,$$

由拉格朗日乘子法得

$$R(\Delta,\theta)=-\sum_{i=1}^{m}\log\big\{1+\lambda_1(\theta)\omega_1(x_{I,i},k,\theta,\Delta)\big\}-\sum_{j=1}^{n}\log\big\{1+\lambda_2(\theta)\omega_2(y_{I,j},k,\theta,\Delta)\big\},$$

其中 $\lambda_j(\theta)$, $j=1,2$，分别由以下方程确定

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\omega_1(x_{I,i},k,\theta,\Delta)}{1+\lambda_1(\theta)\omega_1(x_{I,i},k,\theta,\Delta)}=0, \tag{1}$$

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\omega_2(y_{I,j},k,\theta,\Delta)}{1+\lambda_2(\theta)\omega_2(y_{I,j},k,\theta,\Delta)}=0. \tag{2}$$

令 $\partial R(\theta,\Delta)/\partial\theta=0$，得到经验似然方程

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\alpha_1(x_{I,i},k,\theta,\Delta)}{1+\lambda_1(\theta)\omega_1(x_{I,i},k,\theta,\Delta)}+\frac{1}{m}\sum_{j=1}^{n}\frac{\alpha_2(y_{I,j},k,\theta,\Delta)}{1+\lambda_2(\theta)\omega_2(y_{I,j},k,\theta,\Delta)}=0, \tag{3}$$

其中

$$\alpha_1(x_{I,i},k,\theta,\Delta)=\frac{\partial\omega_1(x_{I,i},k,\theta,\Delta)}{\partial\theta}, \tag{4}$$

$$\alpha_2(y_{I,j},k,\theta,\Delta)=\frac{\partial\omega_2(y_{I,j},k,\theta,\Delta)}{\partial\theta}. \tag{5}$$

用 θ_0 表示 θ 的真值。为了得到本文结果我们需要如下正则条件。

条件1 $\theta_0\in\Omega$，并且 Ω 是一个开区间；

条件2 记 $f(t)=\partial F(t)/\partial t$ 和 $g(t)=\partial G(t)/\partial t$。对某个 $t_0\geq 2$，若 $f^{(t_0-1)}(t)$ 存在，并且在 θ_0 的某个领域内连续，同时 $g^{(t_0-1)}(t)$ 存在，在 $\theta_0+\Delta$ 的某个领域内连续，且 $f(\theta_0)g(\theta_0+\Delta)>0$；

条件3 当 $m,n\to\infty$ 时，有 $\frac{n}{m}\to h\ (0<h<\infty)$；

条件4 对 $i=1,2,K_i$ 有界并满足一阶李普希兹条件，$K_i^{(2)}(\cdot)$ 存在且有界。对常数 $c>0$，有

$$\int_{|u|>c/a^{t_0}}K_1(u)\mathrm{d}u=O(a^{t_0}),\quad \int_{|u|>c/b^{t_0}}K_2(u)\mathrm{d}u=O(b^{t_0}),\quad \int|u^{t_0}K_i(u)|\mathrm{d}u<\infty,$$

$$\int|K_i(u)|\mathrm{d}u<\infty,\qquad \int u^jK_i(u)\mathrm{d}u=\begin{cases}1, & j=0,\\ 0, & 1\leq j\leq t_0-1;\end{cases}$$

条件5 存在$r\,(1/3<r<1/2)$，当$m,\,n\to\infty$时，有$n^r a\to\infty$及$n^r b\to\infty$，$n^{\frac{1}{2}}a^{t_0}\to 0$，$n^{\frac{1}{2}}b^{t_0}\to 0$。

条件6 $\{Z_{x_i}=(x_i,u_i,\delta_{x_i}),\ i=1,\cdots,m\}$和$\{Z_{y_j}=(y_j,v_j,\delta_{y_j}),\ j=1,\cdots,n\}$相互独立。

另外，设数学变量$\gamma\in\mathbf{R}^p$，$\zeta\in\mathbf{R}^q$，记

$$P_1=P(\delta_x=1),\quad S_1=E\Big(\frac{\delta_x uu'}{A^2(u_i)}\Big),\quad P_2=P(\delta_y=1),\quad S_2=E\Big(\frac{\delta_y vv'}{B^2(v_j)}\Big),$$

$$\begin{aligned}H(Z_{x_i},Z_{x_j},\gamma)=\frac{1}{2}\big\{&\delta_{x_i}\delta_{x_j}\omega_1(x_i,\theta_0,\Delta)+(1-\delta_{x_i})\delta_{x_j}\omega_1\big(u_i{}'\gamma+A(u_i)A^{-1}(u_j)(x_j-u_j{}'\gamma),\theta_0,\Delta\big)\\&+\delta_{x_j}\delta_{x_i}\omega_1(x_j,\theta_0,\Delta)+(1-\delta_{x_j})\delta_{x_i}\omega_1\big(u_j{}'\gamma+A(u_j)A^{-1}(u_i)(x_i-u_i{}'\gamma),\theta_0,\Delta\big)\big\},\end{aligned}$$

$$\begin{aligned}G(Z_{y_i},Z_{y_j},\zeta)=\frac{1}{2}\big\{&\delta_{y_i}\delta_{y_j}\omega_2(y_i,\theta_0,\Delta)+(1-\delta_{y_i})\delta_{y_j}\omega_2\big(v_i{}'\zeta+B(v_i)B^{-1}(v_j)(y_j-v_j{}'\zeta),\theta_0,\Delta\big)\\&+\delta_{y_j}\delta_{y_i}\omega_2(y_j,\theta_0,\Delta)+(1-\delta_{y_j})\delta_{y_i}\omega_2\big(v_j{}'\zeta+B(v_j)B^{-1}(v_i)(y_i-v_i{}'\zeta),\theta_0,\Delta\big)\big\},\end{aligned}$$

$$H_1(Z_{x_i},\gamma)=E\big\{H(Z_{x_i},Z_{x_j},\gamma)\mid Z_{x_i}\big\},\quad i\neq j;\qquad h_x(\gamma)=EH(Z_{x_i},Z_{x_j},\gamma),\quad h_{x0}(\gamma)=\frac{\partial h_x(\gamma)}{\partial\gamma},$$

$$G_1(Z_{y_i},\zeta)=E\big\{G(Z_{y_i},Z_{y_j},\zeta)\mid Z_{y_i}\big\},\quad i\neq j;\qquad h_y(\zeta)=EG(Z_{y_i},Z_{y_j},\zeta),\quad h_{y0}(\zeta)=\frac{\partial h_y(\zeta)}{\partial\zeta}.$$

定理1 设条件1至条件5满足，则依概率趋于1，存在经验似然方程(3)的一个根$\theta_{m,n}$，$R(\Delta,\theta)$在$\theta_{m,n}$处达到局部最大值，且当$m,\,n\to\infty$，有

$$\sqrt{m}(\theta_{m,n}-\theta_0)\xrightarrow{d}N\Big(0,\ \frac{1}{c_0^2}\big\{f^2(\theta_0)\sigma_1^2\sigma_4^2+hg^2(\theta_0+\Delta)\sigma_2^2\sigma_3^2\big\}\Big),$$

$$-2R(\Delta,\theta_{m,n})\xrightarrow{d}\frac{hg^2(\theta_0+\Delta)\sigma_1^2+f^2(\theta_0)\sigma_2^2}{c_0}\chi_1^2,$$

其中

$$\begin{aligned}c_0&=f^2(\theta_0)\sigma_4^2+hg^2(\theta_0+\Delta)\sigma_3^2,\\ \sigma_1^2&=P_1^{-2}E\left[2H_1(Z_x,\beta)+h_{x0}(\beta)S_1^{-1}\delta_xA^{-2}(u)u(x-u'\beta)\right]^2+\frac{1}{k}q(1-q)\\&\quad-\frac{1}{k}E\left\{P_1(u)\,E\left[\omega_1^2(x,\theta_0,\Delta)\mid u\right]\right\}-\frac{1}{k}E\left\{(1-P_1(u))\,E^2\left[\omega_1(x,\theta_0,\Delta)\mid u\right]\right\},\\ \sigma_2^2&=P_2^{-2}E\left[2G_1(Z_y,\rho)+h_{y0}(\rho)S_2^{-1}\delta_yB^{-2}(v)v(y-v'\rho)\right]^2+\frac{1}{k}q(1-q)\\&\quad-\frac{1}{k}E\left\{P_2(v)\,E\left[\omega_2^2(y,\theta_0,\Delta)\mid v\right]\right\}-\frac{1}{k}E\left\{(1-P_2(v))\,E^2\left[\omega_2(y,\theta_0,\Delta)\mid v\right]\right\},\\ \sigma_3^2&=\frac{1}{k}q(1-q)+\frac{k-1}{k}E\left\{P_1(u)\,E\left[\omega_1^2(x,\theta_0,\Delta)\mid u\right]\right\}\\&\quad+\frac{k-1}{k}E\left\{(1-P_1(u))\,E^2\left[\omega_1(x,\theta_0,\Delta)\mid u\right]\right\},\\ \sigma_4^2&=\frac{1}{k}q(1-q)+\frac{k-1}{k}E\left\{P_2(v)\,E\left[\omega_2^2(y,\theta_0,\Delta)\mid v\right]\right\}\\&\quad+\frac{k-1}{k}E\left\{(1-P_2(v))\,E^2\left[\omega_2(y,\theta_0,\Delta)\mid v\right]\right\}.\end{aligned}$$

由以上定理知Δ经验似然比统计量的极限分布为加权χ_1^2，为了构造Δ经验似然置信区间，需要给出权

$$a_0(\Delta)=\frac{hg^2(\theta_0+\Delta)\sigma_1^2+f^2(\theta_0)\sigma_2^2}{c_0}$$

的相合估计，我们可用常用的Plug-in进行估计，得到$a_0(\Delta)$的一个相合估计$\hat{a}_0(\Delta)$，由定理1得到Δ(渐近置信水平为$1-\alpha$)的经验似然置信区间$\{\Delta:-2\hat{a}_0^{-1}(\Delta)R(\Delta,\theta_{m,n})\le Z_\alpha\}$，其中$Z_\alpha$满足$P(\chi_1^2\le Z_\alpha)=1-\alpha$。

3 定理1的证明

为证明定理1，我们先给出如下引理。

引理1 在定理1的条件下，当$m,n\to\infty$时，有

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\omega_1(x_{I,i},k,\theta_0,\Delta)\xrightarrow{d}N(0,\sigma_1^2),\quad \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\omega_2(y_{I,j},k,\theta_0,\Delta)\xrightarrow{d}N(0,\sigma_2^2),\tag{6}$$

其中σ_1^2,σ_2^2定义同定理1。

证明 记$\mathcal{B}_m=\sigma\big((x_i,u_i,\delta_{x_i}),\ i=1,\cdots,m\big)$，给定$\mathcal{B}_m$下，由填补过程中随机化所产生的条件概率，条件期望和条件方差分别记为P^*，E^*，Var^*。先证明(6)的第一式。

$$\begin{aligned}&\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\omega_1\big(x_{I,i},k,\theta_0,\Delta\big)\\
&=\sqrt{m}\cdot\frac{1}{m}\sum_{i\in s_{r_x}}\omega_1\big(x_i,\theta_0,\Delta\big)+\sqrt{m}\cdot\frac{1}{m}\sum_{i\in s_{m_x}}\Big\{\frac{1}{k}\sum_{l=1}^{k}\omega_1\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta,\Delta\big)\Big\}\\
&=\xi_m+\eta_m+o_p(1),\end{aligned}$$

其中$\xi_m=P_1^{-1}\sqrt{m}V_{xm}(\hat{\beta}_r)$，且

$$\begin{aligned}V_{xm}(\hat{\beta}_r)=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\Big\{&\delta_{x_i}\delta_{x_j}\omega_1\big(x_i,\,\theta_0,\,\Delta\big)\\
&+(1-\delta_{x_i})\delta_{x_j}\omega_1\big(u_i{}'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-u_j{}'\hat{\beta}_r),\,\theta_0,\,\Delta\big)\Big\},\end{aligned}$$

$$\begin{aligned}\eta_m=\frac{1}{k\sqrt{m}}\sum_{i\in s_{m_x}}\sum_{l=1}^{k}\Big\{&\omega_1\big(u_i{}'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)\\
&-\frac{1}{r_x}\sum_{j\in s_{r_x}}\omega_1\big(u_i{}'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-u_j{}'\hat{\beta}_r),\,\theta_0,\,\Delta\big)\Big\}.\end{aligned}$$

我们将利用文献[13]中的定理2来证明(6)的第一式。为此，需要分别求出ξ_m和η_m的极限分

布。首先考虑 $V_{xm}(\hat{\beta}_r)$。将 $V_{xm}(\gamma)$ 对称化得

$$
\begin{aligned}
V_{xm}(\gamma) &= \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\{\,\delta_{x_i}\delta_{x_j}\omega_1(x_i,\,\theta_0,\,\Delta)\\
&\quad +(1-\delta_{x_i})\delta_{x_j}\omega_1\big(u_i{}'\gamma+A(u_i)A^{-1}(u_j)(x_j-u_j{}'\gamma),\,\theta_0,\,\Delta\big)\\
&\quad +\delta_{x_j}\delta_{x_i}\omega_1(x_j,\theta_0,\Delta)+(1-\delta_{x_j})\delta_{x_i}\omega_1\big(u_j{}'\gamma+A(u_j)A^{-1}(u_i)x_i-u_i{}'\gamma,\theta_0,\Delta\big)\}\\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}H(Z_{x_i},\,Z_{x_j},\,\gamma).
\end{aligned}
$$

显然，$V_{xm}(\gamma)$ 是一个 V 统计量，$V_{xm}(\hat{\beta}_r)$ 是含参数估计值的 V 统计量，$V_{xm}(\hat{\beta}_r)$ 相应的 U 统计量为

$$
U_{xm}(\hat{\beta}_r)=\frac{2}{m(m-1)}\sum_{1\le i<j\le m}H(Z_{x_i},\,Z_{x_j},\,\hat{\beta}_r),
$$

且由文献 [15] 或文献 [16] 知 $\sqrt{m}\,[\,V_{xm}(\hat{\beta}_r)-U_{xm}(\hat{\beta}_r)\,]=o_p(1)$，故

$$
\sqrt{m}\,V_{xm}(\hat{\beta}_r)=\sqrt{m}\,[\,U_{xm}(\hat{\beta}_r)-h_x(\beta)\,]+\sqrt{m}\,h_x(\beta)+o_p(1). \tag{7}
$$

由文献 [10] 知 $\sqrt{m}\,[\,U_{xm}(\hat{\beta}_r)-h_x(\beta)\,]$ 与

$$
\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\{\,2H_1(Z_{x_i},\beta)+h_{x0}(\beta)S_1^{-1}\delta_{x_i}A^{-2}(u_i)u_i(x_i-u_i{}'\beta)\,\} \tag{8}
$$

有相同的极限分布，其中 $H_1(Z_{x_i},\beta)=E\,[\,H(Z_{x_i},Z_{x_j},\beta)\,|\,Z_{x_i}\,]$, $i\neq j$。下面先证明 $\sqrt{m}\,h_x(\beta)=o_p(1)$。由 $H(Z_{x_i},Z_{x_j},\gamma)$ 的形式知 $h_x(\gamma)=E\,[\,\delta_{x_1}\delta_{x_2}\omega_1(x_1,\theta_0,\Delta)\,]+a_x(\gamma)$，其中

$$
a_x(\gamma)=E\,\big[\,(1-\delta_{x_1})\delta_{x_2}\omega_1\big(u_1{}'\gamma+A(u_1)A^{-1}(u_2)(x_2-u_2{}'\gamma),\theta_0,\Delta\big)\,\big].
$$

由 MAR 假定知

$$
\begin{aligned}
E\{\,\delta_{x_1}\delta_{x_2}\omega_1(x,\,\theta_0,\,\Delta)\,\}&=P_1E\{\,E(\delta_{x_1}\,|\,u_1)E\,[\,\omega_1(x,\theta_0,\Delta)\,|\,u_1\,]\,\}\\
&=P_1E\{\,P_1(u)E\,[\,\omega_1(x,\theta_0,\Delta)\,|\,u\,]\,\},
\end{aligned} \tag{9}
$$

$$
a_x(\gamma)=E\{(1-P_1(u_1))P_1(u_2)E\,[\,\omega_1(u_1{}'\gamma+A(u_1)A^{-1}(u_2)(x_2-u_2{}'\gamma),\theta_0,\Delta)\,|\,u_1,u_2\,]\},
$$

因此，在 $\gamma=\beta$ 时，有

$$
\begin{aligned}
a_x(\beta) &= E\{\,(1-P_1(u_1))P_1(u_2)E\,[\,\omega_1(u_1{}'\beta+A(u_1)\varepsilon_2,\theta_0,\Delta)\mid u_1,u_2\,]\}\\
&= E\{\,(1-P_1(u_1))P_1(u_2)E\,[\,\omega_1(u_1{}'\beta+A(u_1)\varepsilon_1,\theta_0,\Delta)\mid u_1,u_2\,]\,\}\\
&= P_1E\{\,(1-P_1(u_1))E\,[\,\omega_1(x,\theta_0,\Delta)\mid u_1\,]\,\}\\
&= P_1E\omega_1(x,\theta_0,\Delta)-P_1E\{\,P_1(u)E\,[\,\omega_1(x,\theta_0,\Delta)\mid u_1\,]\,\}.
\end{aligned} \tag{10}
$$

由 (9) 和 (10) 知

$$
h_x(\beta)=P_1E\{\,P_1(u)E\,[\,\omega_1(x,\theta_0,\Delta)\mid u\,]\,\}+a_x(\beta)=P_1E\omega_1(x,\theta_0,\Delta).
$$

从文献[12]可知$E\omega_1(x_i,\theta_0,\Delta)=O(a^{t_0})$，因此，由条件5有$\sqrt{m}\,h_x(\beta)=O(m^{\frac{1}{2}}a^{t_0})=o(1)$。另外MAR假定下知$EH_1(Z_{x_i},\beta)=h_x(\beta)=o(1)$，且$E\{h_{x0}(\beta)S_1^{-1}\delta_{x_i}A^{-2}(u_i)u_i(x_i-u_i'\beta)\}=0$，于是由(7),(8)和中心极限定理知

$$\sigma_{1m}^{-2}\xi_m \xrightarrow{d} N(0,1), \tag{11}$$

其中$\sigma_{1m}^2=E\{2H_1(Z_x,\beta)+h_{x0}(\beta)S_1^{-1}\delta_{x_i}A^{-2}(u)u(x-u'\beta)\}^2$。

下面考虑η_m。在MAR假定下有

$$E^*\omega_1\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)=\frac{1}{r_x}\sum_{j\in s_{r_x}}\omega_1\big(u_i'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-u_j'\hat{\beta}_r),\theta_0,\Delta\big).$$

故$E^*\eta_m=0$，且

$$\begin{aligned}\mathrm{Var}^*\eta_m &= \mathrm{Var}\Big\{\frac{1}{k\sqrt{m}}\sum_{i\in s_{m_x}}\sum_{l=1}^{k}\omega_1\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)-E^*\omega_1\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)\Big\}\\ &= \frac{1}{km}\sum_{i\in s_{m_x}}\big\{E^*\omega_1^2\big(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)-\big[E^*\omega_1(u_i'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\theta_0,\Delta)\big]^2\big\}\\ &= I_{xm1}-I_{xm2},\end{aligned} \tag{12}$$

其中

$$I_{xm1}=\frac{1}{kmr_x}\sum_{i\in s_{m_x}}\sum_{j\in s_{r_x}}\omega_1^2\big(u_i'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-u_j'\hat{\beta}_r),\ \theta_0,\ \Delta\big),$$

$$I_{xm2}=\frac{1}{km}\sum_{i\in s_{m_x}}\Big\{\frac{1}{r_x}\sum_{j\in s_{r_x}}\omega_1\big(u_i'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-u_j'\hat{\beta}_r),\ \theta_0,\ \Delta\big)\Big\}^2.$$

下面分别导出I_{xm1}和I_{xm2}依概率收敛意义下的极限。

$$\begin{aligned}I_{xm1} &= \frac{1}{kmr_x}\sum_{i=1}^{m}\sum_{j=1}^{m}(1-\delta_{x_i})\delta_{x_j}\omega_1^2\big(u_i'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-u_j'\hat{\beta}_r),\theta_0,\Delta\big)\\ &= \frac{1}{k}\cdot P_1^{-1}\cdot\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}H_2(Z_{x_i},Z_{x_j},\hat{\beta}_r)+o_p(1),\end{aligned}$$

其中

$$\begin{aligned}H_2(Z_{x_i},Z_{x_j},\hat{\beta}_r) &= \frac{1}{2}\{(1-\delta_{x_i})\delta_{x_j}\omega_1^2\big(u_i'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-u_j'\hat{\beta}_r),\theta_0,\Delta\big)\\ &\quad+\delta_{x_i}(1-\delta_{x_j})\omega_1^2\big(u_j'\hat{\beta}_r+A(u_j)A^{-1}(u_i)(x_i-u_i'\hat{\beta}_r),\theta_0,\Delta\big)\}.\end{aligned}$$

则$H_2(Z_{x_i},Z_{x_j},\hat{\beta}_r)$是一个带估计参数的$V$统计量，由文献[15]或文献[16]知

$$\begin{aligned}&\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}H_2(Z_{x_i},Z_{x_j},\hat{\beta}_r)\\ =\ & EH_2(Z_{x_i},Z_{x_j},\beta_r)+o_p(1)\\ =\ & E\big\{P_1(u_2)(1-P_1(u_1))E\big[\omega_1^2(u_1'\beta+A(u_1)A^{-1}(u_2)(x_2-u_2'\beta),\theta_0,\Delta)\mid u_1,u_2\big]\big\}+o_p(1)\\ =\ & P_1E\big\{(1-P_1(u))E\big[\omega_1^2(x,\theta_0,\Delta)\mid u\big]\big\}+o_p(1).\end{aligned}$$

故

$$I_{xm1} = \frac{1}{k}E\left\{ (1 - P_1(u))E\left[\omega_1^2(x, \theta_0, \Delta) \mid u\right] \right\} + o_p(1). \tag{13}$$

类似于 (10) 的推导得

$$\begin{aligned}
I_{xm2} &= \frac{1}{k} \cdot \left\{P_1^{-2} + o_p(1)\right\} \frac{1}{m^3} \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}(1 - \delta_{x_i})\delta_{x_j}\delta_{x_k} \\
&\quad \omega_1\big(u_i{}'\hat{\beta}_r + A(u_i)A^{-1}(u_j)(x_j - u_j{}'\hat{\beta}_r), \theta_0, \Delta\big) \\
&\quad \times \omega_1\big(u_i{}'\hat{\beta}_r + A(u_i)A^{-1}(u_k)(x_k - u_k{}'\hat{\beta}_r), \theta_0, \Delta\big) \\
&= \frac{1}{k} \cdot P_1^{-2}\, E\left\{E\left[(1 - \delta_{x_1})\delta_{x_2}\delta_{x_3}\omega_1(u_1{}'\beta + A(u_1)A^{-1}(u_2)(x_2 - u_2{}'\beta), \theta_0, \Delta)\right.\right. \\
&\quad \left.\left. \times\, \omega_1(u_1{}'\beta + A(u_1)A^{-1}(u_3)(x_3 - u_3{}'\beta), \theta_0, \Delta) \mid u_1, u_2, u_3 \right] \right\} + o_p(1) \\
&= \frac{1}{k} \cdot E\left\{ (1 - P_1(u))E^2\left[\omega_1(x, \theta_0, \Delta) \mid u\right] \right\} + o_p(1).
\end{aligned} \tag{14}$$

由文献 [12] 知 $E\omega_1^2(x, \theta_0, \Delta) = q(1 - q) + o(1)$，因此，结合 (12), (13), (14) 知

$$\begin{aligned}
\sigma_{2m}^2 &= \mathrm{Var}^*\eta_m \\
&= \frac{q(1-q)}{k} - \frac{E\{P_1(u)E[\omega_1^2(x_1, \theta_0, \Delta)|u]\}}{k} - \frac{E\{(1 - P_1(u))E^2[\omega_1(x, \theta_0, \Delta)|u]\}}{k} + o_p(1).
\end{aligned}$$

故由独立随机变量 Berry-Esseen's 中心极限定理有

$$\sup_t \left|P^*(\sigma_{2m}^{-1}\eta_m \le t) - \Phi(t)\right| = o_{p^*}(1), \tag{15}$$

于是，由文献 [13] 中定理 2 和 (11), (15) 知

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \omega_1(x_{I,i}, k, \theta_0, \Delta) \xrightarrow{d} N(0, \sigma_1^2), \tag{16}$$

即 (6) 的第一式得证。同理可证 (6) 的第二式成立。

引理 2 在定理 1 的条件下，当 $m, n \to \infty$ 时，有

$$\frac{1}{m} \sum_{i=1}^{m} \omega_1^2(x_{I,i}, k, \theta_0, \Delta) = \sigma_3^2 + o_p(1), \quad \frac{1}{n} \sum_{j=1}^{n} \omega_2^2(y_{I,j}, k, \theta_0, \Delta) = \sigma_4^2 + o_p(1), \tag{17}$$

其中 σ_3^2, σ_4^2 的定义同定理 1。

证明 先证 (17) 的第一式。令

$$\frac{1}{m} \sum_{i=1}^{m} \omega_1^2(x_{I,i}, k, \theta_0, \Delta) = I_{xm3} + I_{xm4},$$

其中

$$I_{xm3} = \frac{1}{m} \sum_{i \in s_{r_x}} \omega_1^2(x_i, \theta_0, \Delta), \quad I_{xm4} = \frac{1}{m} \sum_{i \in s_{m_x}} \left\{ \frac{1}{k} \sum_{l=1}^{k} \omega_1^2\big(u_i{}'\hat{\beta}_r + A(u_i)\varepsilon_{il}^*, \theta_0, \Delta\big) \right\}.$$

由MAR假定知

$$
\begin{aligned}
I_{xm3} &= \frac{1}{m}\sum_{i=1}^{m}\{\delta_{x_i}\omega_1^2(x_i,\,\theta_0,\,\Delta)\} \\
&= E\{P_1(u)E[\omega_1^2(x_1,\,\theta_0,\,\Delta)\mid u]\} + o_p(1).
\end{aligned}
\tag{18}
$$

下面求 I_{xm4} 在依概率收敛意义下的极限。由于

$$
\begin{aligned}
& E^*\omega_1^2\big(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big) \\
={}& \frac{1}{r_x}\sum_{j\in s_{r_x}}\omega_1^2\big(u_i{}'\hat\beta_r + A(u_i)A^{-1}(u_j)(x_j - u_j{}'\hat\beta_r),\theta_0,\Delta\big),
\end{aligned}
$$

故

$$
\begin{aligned}
& E^*\Big\{\frac{1}{k}\sum_{l=1}^{k}\omega_1\big(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)\Big\}^2 \\
={}& \frac{1}{k^2}E^*\Big\{\sum_{l=1}^{k}\big[\omega_1(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta) - E^*\omega_1(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta)\big] \\
& + kE^*\omega_1\big(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)\Big\}^2 \\
={}& \frac{1}{k}\cdot\frac{1}{r_x}\sum_{j\in s_{r_x}}\omega_1^2\big(u_i{}'\hat\beta_r + A(u_i)A^{-1}(u_j)(x_j - u_j{}'\hat\beta_r),\theta_0,\Delta\big) \\
& + \frac{k-1}{k}\Big\{\frac{1}{r_x}\sum_{j\in s_{r_x}}\omega_1\big(u_i{}'\hat\beta_r + A(u_i)A^{-1}(u_j)(x_j - u_j{}'\hat\beta_r),\theta_0,\Delta\big)\Big\}^2.
\end{aligned}
$$

令

$$
\begin{aligned}
J_m ={}& \frac{1}{m}\sum_{i\in s_{m_x}}\Big\{\Big[\frac{1}{k}\sum_{l=1}^{k}\omega_1\big(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)\Big]^2 \\
& - \frac{1}{kr_x}\sum_{j\in s_{r_x}}\omega_1^2\big(u_i{}'\hat\beta_r + A(u_i)A^{-1}(u_j)(x_j - u_j{}'\hat\beta_r),\theta_0,\Delta\big) \\
& - \frac{k-1}{k}\Big\{\frac{1}{r_x}\sum_{j\in s_{r_x}}\omega_1\big(u_i{}'\hat\beta_r + A(u_i)A^{-1}(u_j)(x_j - u_j{}'\hat\beta_r),\theta_0,\Delta\big)\Big\}^2\Big\} \\
={}& \frac{m_x}{m}\cdot\frac{1}{m_x}\sum_{i\in s_{m_x}}\Big\{\Big[\frac{1}{k}\sum_{l=1}^{k}\omega_1\big(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)\Big]^2 \\
& - E^*\Big[\frac{1}{k}\sum_{l=1}^{k}\omega_1\big(u_i{}'\hat\beta_r + A(u_i)\varepsilon_{il}^*,\theta_0,\Delta\big)\Big]^2\Big\},
\end{aligned}
$$

在MAR假定下，由大数定律知 $J_m = \{1 - P_1 + o_p(1)\} \cdot o_{p^*}(1) = o_p(1)$，故

$$
\begin{aligned}
I_{xm4} &= \frac{1}{kmr_x} \sum_{i\in s_{m_x}} \sum_{j\in s_{r_x}} \omega_1^2\big(u_i'\hat{\beta}_r + A(u_i)A^{-1}(u_j)(x_j - u_j'\hat{\beta}_r), \theta_0, \Delta\big) \\
&\quad + \frac{k-1}{k} \cdot \frac{1}{m} \sum_{i\in s_{m_x}} \Big\{\frac{1}{r_x} \sum_{j\in s_{r_x}} \omega_1\big(u_i'\hat{\beta}_r + A(u_i)A^{-1}(u_j)(x_j - u_j'\hat{\beta}_r), \theta_0, \Delta\big)\Big\}^2 + o_p(1) \\
&= I_{xm1} + (k-1)I_{xm2} + o_p(1),
\end{aligned}
$$

其中 I_{xm1}，I_{xm2} 的定义同引理1。于是由引理1的证明知

$$
\begin{aligned}
I_{xm4} &= \frac{1}{k} E\big\{(1 - P_1(u))E\,[\omega_1^2(x, \theta_0, \Delta) \mid u]\big\} \\
&\quad + \frac{k-1}{k} E\big\{(1 - P_1(u))E^2\,[\omega_1(x, \theta_0, \Delta) \mid u]\big\} + o_p(1).
\end{aligned} \tag{19}
$$

故由(18)和(19)知(17)的第一式成立。同理可证得(17)的第二式成立，引理2得证。

引理3　若 $1/3 < \eta < 1/2$ 并且定理1的条件满足，则对一切 $\theta \in \{\theta : |\theta - \theta_0| \le cn^{-\eta}\}$ 有

$$
\frac{1}{m}\sum_{i=1}^{m} \alpha_1(x_{I,i}, k, \theta, \Delta) = f(\theta_0) + o_p(1), \tag{20}
$$

$$
\frac{1}{n}\sum_{j=1}^{n} \alpha_2(y_{I,j}, k, \theta, \Delta) = g(\theta_0 + \Delta) + o_p(1). \tag{21}
$$

证明　先证(20)。记

$$
f_{xm1}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}\alpha_1(x_i, \theta, \Delta),
$$

$$
f_{xm2}(\theta) = \frac{1}{m}\sum_{i=1}^{m}(1 - \delta_{x_i})\Big\{\frac{1}{k}\sum_{l=1}^{k} \alpha_1\big(u_i'\hat{\beta}_r + A(u_i)\varepsilon_{il}^*, \theta, \Delta\big)\Big\},
$$

则由(3)知

$$
\frac{1}{m}\sum_{i=1}^{m} \alpha_1(x_{I,i}, k, \theta, \Delta) = f_{xm1}(\theta) + f_{xm2}(\theta).
$$

由MAR假定和条件4，条件5知

$$
\begin{aligned}
f_{xm1}(\theta) &= \frac{1}{ma}\sum_{i\in s_{r_x}} K_1\Big(\frac{\theta_0 - x_i}{a}\Big) + \frac{1}{ma}\sum_{i\in s_{r_x}} \Big[K_1'\Big(\frac{\theta^* - x_i}{a}\Big)(\theta - \theta_0)\Big] \\
&= \frac{1}{ma}\sum_{i\in s_{r_x}} K_1\Big(\frac{\theta_0 - x_i}{a}\Big) + O_p(a^{-1}n^{-\eta}) = E\delta_x\alpha_1(x, \theta_0, \Delta) + o_p(1),
\end{aligned}
$$

其中 θ^* 介于 θ 和 θ_0 之间。故

$$
f_{xm1}(\theta) = E\big\{P_1(u)E\,[\alpha_1(x, \theta_0, \Delta) \mid u]\big\} + o_p(1). \tag{22}
$$

同理，有

$$f_{xm2}(\theta)=\frac{1}{m}\sum_{i=1}^{m}(1-\delta_{x_i})\Big\{\frac{1}{k}\sum_{l=1}^{k}\alpha_1\big({u_i}'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\,\theta_0,\,\Delta\big)\Big\}+o_p(1).$$

由弱大数定律知

$$\frac{1}{m}\sum_{i=1}^{m}\Big\{(1-\delta_{x_i})\Big[\frac{1}{k}\sum_{l=1}^{k}\alpha_1\big({u_i}'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\,\theta_0,\,\Delta\big)\Big]$$
$$-E^*(1-\delta_{x_i})\Big[\frac{1}{k}\sum_{l=1}^{k}\alpha_1\big({u_i}'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\,\theta_0,\,\Delta\big)\Big]\Big\}=o_{p^*}(1).$$

因此

$$\frac{1}{m}\sum_{i=1}^{m}(1-\delta_{x_i})\Big\{\frac{1}{k}\sum_{l=1}^{k}\alpha_1\big({u_i}'\hat{\beta}_r+A(u_i)\varepsilon_{il}^*,\,\theta_0,\,\Delta\big)\Big\}=P_1^{-1}\cdot\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}H_3(Z_{x_i},Z_{x_j},\hat{\beta}_r)+o_p(1),$$

其中

$$H_3(Z_{x_i},Z_{x_j},\hat{\beta}_r)=\frac{1}{2}\big[(1-\delta_{x_i})\delta_{x_j}\alpha_1({u_i}'\hat{\beta}_r+A(u_i)A^{-1}(u_j)(x_j-{u_j}'\hat{\beta}_r),\theta_0,\Delta)$$
$$+\delta_{x_i}(1-\delta_{x_j})\alpha_1({u_j}'\hat{\beta}_r+A(u_j)A^{-1}(u_i)(x_i-{u_i}'\hat{\beta}_r),\theta_0,\Delta)\big].$$

由文献 [15] 或 [16] 知

$$\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}H_3(Z_{x_i},Z_{x_j},\hat{\beta}_r)=P_1E\big\{(1-P_1(u))E\,[\,\alpha_1(x,\,\theta_0,\,\Delta)\mid u\,]\big\}+o_p(1).$$

从而

$$f_{xm2}(\theta)=E\big\{(1-P_1(u))E\,[\,\alpha_1(x,\,\theta_0,\,\Delta)\mid u\,]\big\}+o_p(1).\tag{23}$$

由 (22), (23) 知

$$\frac{1}{m}\sum_{i=1}^{m}\alpha_1(x_{I,i},k,\theta_0,\Delta)=E\alpha_1(x,\theta_0,\Delta)+o_p(1)=\frac{1}{a}EK_1\Big(\frac{\theta_0-x}{a}\Big)+o_p(1).$$

由文献 [14] 中定理 1 和条件 4 有

$$\frac{1}{a}EK_1\Big(\frac{\theta_0-x}{a}\Big)=\lim_{m\to\infty}\frac{1}{a}\int_{-\infty}^{+\infty}K_1\Big(\frac{\theta_0-x}{a}\Big)f(x)\mathrm{d}x=f(\theta_0).$$

进一步有

$$\frac{1}{m}\sum_{i=1}^{m}\alpha_1(x_{I,i},k,\theta,\Delta)=f(\theta_0)+o_p(1).$$

即 (20) 式成立。同理可证 (21) 式。

结合上述引理 1-3，类似于文献 [12]，我们容易得到下述引理 4-6。

引理 4 若 $1/3<\eta<1/2$ 且定理 1 的条件满足，则当 $m,\,n\to\infty$ 时，有

$$\lambda_1(\theta)=O_p(n^{-\eta}),\quad \lambda_2(\theta)=O_p(n^{-\eta})$$

对一切 $\theta \in \{\theta: |\theta-\theta_0| \le cn^{-\eta}\}$ 成立，其中 c 是正常数。

引理 5 若 $1/3 < \eta < 1/2$ 且定理 1 的条件满足，则依概率 1 存在方程 (3) 的一个根 $\theta_{m,n}$，当 $m, n \to \infty$ 时，有 $|\theta_{m,n}-\theta_0| = O_p(n^{-\eta})$，且 $R(\Delta,\theta)$ 在 $\theta_{m,n}$ 处达到它的局部最大值。

引理 6 若定理 1 的条件满足，$\theta_{m,n}$ 如引理 5 中所定义，则当 $m,n\to\infty$ 时，有

$$\sqrt{m}\,(\theta_{m,n}-\theta_0) \xrightarrow{d} N\Big(0, \frac{1}{c_0^2}\{f^2(\theta_0)\sigma_1^2\sigma_4^2 + hg^2(\theta_0+\Delta)\sigma_2^2\sigma_3^2\}\Big),$$

$$\lambda_1(\theta_{m,n}) = -\frac{hg(\theta_0+\Delta)}{f(\theta_0)}\lambda_2(\theta_{m,n}) + o_p(1), \quad \sqrt{m}\,\lambda_2(\theta_{m,n}) \xrightarrow{d} N(0,\sigma^2),$$

其中

$$\sigma^2 = \frac{f^2(\theta_0)}{c_0^2}\{g^2(\theta_0+\Delta)\sigma_1^2 + h^{-1}f^2(\theta_0)\sigma_2^2\},$$

c_0, σ_1^2, σ_2^2, σ_3^2, σ_4^2 同定理 1。

定理 1 的证明 记

$$R_1(\Delta,\theta) = -\sum_{i=1}^m \log\{1+\lambda_1(\theta)\omega_1(x_{I,i},\, k,\, \theta,\, \Delta)\},$$

$$R_2(\Delta,\theta) = -\sum_{j=1}^n \log\{1+\lambda_2(\theta)\omega_2(y_{I,j},\, k,\, \theta,\, \Delta)\}.$$

类似于文献 [3] 中定理 1 的证明，由 Taylor 展开知

$$\begin{aligned}-R_1(\Delta,\theta_{m,n}) \;=\; & \sum_{i=1}^m \lambda_{E_1}\omega_1(x_{I,i},k,\theta_{m,n},\Delta)\\ & -\frac{1}{2}\sum_{i=1}^m \lambda_{E_1}^2\omega_1^2(x_{I,i},k,\theta_{m,n},\Delta) + \sum_{i=1}^m O_p\{\lambda_{E_1}^3\omega_1^3(x_{I,i},k,\theta_{m,n},\Delta)\}.\end{aligned}$$

记 $\lambda_{E_1} = \lambda_1(\theta_{m,n})$，$\lambda_{E_2} = \lambda_2(\theta_{m,n})$，则

$$\lambda_{E_1} = \frac{1}{m}\sum_{i=1}^m \omega_1(x_{I,i},k,\theta_{m,n},\Delta)\cdot\Big\{\frac{1}{m}\sum_{i=1}^m \omega_1^2(x_{I,i},k,\theta_{m,n},\Delta)\Big\}^{-1} + O_p(\lambda_{E_1}^2).$$

故

$$\frac{1}{m}\sum_{i=1}^m \omega_1(x_{I,i},k,\theta_{m,n},\Delta) = \lambda_{E_1}\cdot\frac{1}{m}\sum_{i=1}^m \omega_1^2(x_{I,i},k,\theta_{m,n},\Delta) + O_p(n^{-2\eta}),$$

因此

$$-2R_1(\Delta,\theta_{m,n}) = m\lambda_{E_1}^2\cdot\frac{1}{m}\sum_{i=1}^m \omega_1^2(x_{I,i},k,\theta_{m,n},\Delta) + o_p(1),$$

同理

$$-2R_2(\Delta,\theta_{m,n}) = n\lambda_{E_2}^2\cdot\frac{1}{n}\sum_{j=1}^n \omega_2^2(y_{I,j},k,\theta_{m,n},\Delta) + o_p(1).$$

故

$$-2R(\Delta,\theta_{m,n}) = m\lambda_{E_1}^2\cdot\frac{1}{m}\sum_{i=1}^{m}\omega_1^2(x_{I,i},k,\theta_{m,n},\Delta)$$

$$+n\lambda_{E_2}^2\cdot\frac{1}{n}\sum_{i=1}^{m}\omega_2^2(y_{I,j},k,\theta_{m,n},\Delta)+o_p(1)=a_0(\Delta)\Big(\frac{\sqrt{m}}{\sigma}\lambda_{E_2}\Big)^2+o_p(1).$$

故由引理6知定理1成立。

4 模拟结果

本节将通过数值模拟研究 $\Delta = G^{-1}(q) - F^{-1}(q)$ 的经验似然置信区间的性质，为此选取以下两线性模型。

$$x = u'\beta+\varepsilon,\quad \beta=4,\quad u\sim N(1,1),\quad \varepsilon\sim N(0,1),$$

$$y = v'\rho+\tau,\quad \rho=2,\quad v\sim N(1,2),\quad \tau\sim N(0,1).$$

考虑下述两种响应概率 (MAR)。

情形1

$$P_1(u)=\begin{cases}0.8+0.2|u-1|, & 若\ |u-1|\le 4,\\ 0.95, & 其它,\end{cases}$$

$$P_2(v)=\begin{cases}0.8+0.2|v-1|, & 若\ |v-1|\le 4\ ,\\ 0.95, & 其它.\end{cases}$$

情形2 $P_1(u)=P_2(v)=0.8$。

对上述两种情形随机产生1000次不完全样本 $\{(x_i,\ u_i,\ \delta_{x_i}),\ (y_j,\ v_j,\ \delta_{y_j}),\ i=1,\cdots,m,\ j=1,\cdots,n\}$，其中 m, n 表示两样本容量。取分位数 $q=0.7$，Δ 的区间估计的置信水平 $1-\alpha=0.95$，窗宽

$$a=\frac{3}{2}m^{-\frac{1}{3}},\quad b=\frac{3}{2}n^{-\frac{1}{3}},$$

核函数

$$K_1(u)=K_2(u)=\frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{u^2}{2}\Big\}.$$

在模拟过程中对定理1中所得的权 $a_0(\Delta)$ 采用常用的Plug-in方法进行估计。用AL表示模拟所得区间的平均长度，CP表示模拟所得区间覆盖 Δ 的真值的平均覆盖率。

从表1和表2的模拟结果可看到如下事实：

1) 当样本容量适当大时，在分数填补法下得到的置信区间平均覆盖率接近置信水平0.95；且在相同的响应概率和相同的填补次数下，当样本容量逐渐增大时，所得区间的平均长度逐渐缩短，区间覆盖真值的平均覆盖率逐渐逼近0.95；

2) 在样本容量和响应概率不变时，随着 k 的增大，所得区间的平均长度呈递减趋势，且区间平均覆盖率与置信水平0.95的接近程度越高；特别地当 $k=1$ 时，即随机填补法下所得区间的平均长度比分数填补 $(k\ge 2)$ 下所得区间平均长度大，模拟结果较差；当 $k\ge 7$ 时，模拟结果近似于 $k=6$ 的情形，结果趋于稳定。

表1: 情形1下的模拟

(m,n)	k	AL	CP(%)
(100,100)	1	0.66999	93.0
	2	0.57249	93.7
	3	0.58325	94.0
	4	0.56109	95.9
	5	0.53732	95.4
	6	0.48930	95.2
(200,200)	1	0.65225	93.5
	2	0.61170	94.3
	3	0.57614	94.7
	4	0.55006	93.9
	5	0.56010	94.6
	6	0.47459	95.2
(300,300)	1	0.64623	93.6
	2	0.60531	93.8
	3	0.57102	94.8
	4	0.54116	94.7
	5	0.50134	95.4
	6	0.48017	95.1

表2: 情形2下的模拟

(m,n)	k	AL	CP(%)
(100,100)	1	1.54247	92.5
	2	1.42450	93.0
	3	1.32682	94.1
	4	1.25730	94.6
	5	1.15483	95.7
	6	1.13761	95.4
(200,200)	1	1.47761	94.0
	2	1.37892	93.2
	3	1.39539	94.8
	4	1.27276	94.7
	5	1.22792	94.5
	6	1.17689	95.2
(300,300)	1	1.43726	92.8
	2	1.41044	93.6
	3	1.34658	95.2
	4	1.31480	94.3
	5	1.15833	95.6
	6	1.09978	95.3

注：填补次数k取为正整数，目前还没有一个统一的方法决定k，也没有最优的k的选取方法(一般讲，k越大，填补方差越小，故最优的k一般不存在)，但可用经验方法大致确定k(见文献[10,11])，k的经验选取原则是：对$k=1,2,\cdots,10$，通过模拟来比较结果的好坏，即选定模拟结果比较稳定(即k再增加时，得到的结果改变不大)时的k，已有的模拟结果表明，k在5-10之间时的结果比较好。

致谢：衷心感谢审稿人对本文提出的宝贵修改意见。

参考文献：

[1] Owen A B. Empirical likelihood ratio confidence intervals for a single functional[J]. Biometrika, 1988, 75: 237-249

[2] 王启华. 经验似然统计推断方法发展综述[J]. 数学进展，2004, 2: 141-150
Wang Q H. Empirical likelihood statistical inference approach: a survey[J]. Advances in Mathematics, 2004, 2: 141-150

[3] Owen A B. Empirical likelihood ratio confidence regions[J]. The Annals of Statistics, 1990, 18: 90-120

[4] 秦永松. 部分线性模型参数的经验似然比置信域[J]. 应用概率统计，1999, 15(4): 363-369
Qin Y S. Empirical likelihood ratio confidence regions in a partly linear model[J]. Chinese Journal of Applied Probability and Statistics, 1999, 15(4): 363-369

[5] 石坚. 高维线性模型中的经验似然[J]. 系统科学与数学，2007, 2: 124-133
Shi J. Empirical likelihood for higher dimensional linear models[J]. Journal of Systems Science and Mathematics, 2007, 2: 124-133

[6] 薛留根，廖靖宇. 删失数据下一类回归模型的参数估计[J]. 工程数学学报，2005, 22(4): 712-718

Xue L G, Liao J Y. Parameter estimation of a regression model under censored data[J]. Chinese Journal of Engineering Mathematics, 2005, 22(4): 712-718
[7] Wang Q H, *et al.* Semiparametric regression analysis with missing response at random[J]. Journal of the American Statistical Association, 2004, 99: 334-345
[8] 王立春. 删失下的指数分布的贝叶斯估计[J]. 工程数学学报，2006, 23(3): 553-558
Wang L C. Bayes estimator for the exponential distribution under censorship[J]. Chinese Journal of Engineering Mathematics, 2006, 23(3): 553-558
[9] Wang Q H, Rao J N K. Empirical likelihood for linear regression models under imputation for missing response[J]. The Canadian Journal Statistics, 2001, 29: 597-608
[10] Qin Y, *et al.* Confidence intervals for marginal parameters under fractional linear regression imputation for missing data[J]. Journal of Multivariate Analysis, 2008, 99: 1232-1259
[11] Kim J K, Fuller W. Fractional hot deck imputation[J]. Biometrika, 2004, 91: 559-578
[12] 秦永松，赵林城. 两总体分位数差异的经验似然比置信区间[J]. 数学年刊，1997, 18A: 687-694
Qin Y S, Zhao L C. Empirical likelihood ratio confidence intervals for the quantile differences of two populations[J]. Chinese Annals of Mathematics, 1997, 18A: 687-694
[13] Chen J, Rao J N K. Asymptotic normality under two-phase sampling designs[J]. Ststistica Sinica, 2007, 17: 1047-1064
[14] Parzen E. On estimation of a probability density function and model[J]. The Annals of Mathematical Statistics, 1962, 3: 1065-1076
[15] Serfling R J. Approximation Theorems of Mathematical Statistics[M]. New York: John Wiley & Sons, 1980
[16] Randles R H. On the asymptotic normality of statistics with estimated parameters[J]. The Annals of Statistics, 1982, 10: 462-474

Empirical Likelihood Confidence Intervals for Quantile Differences of Response Variables in Two Linear Regression Models with Missing Data

WANG Li-rong[1], QIN Yong-song[2], BAI Yun-xia[3], LI Ling[4]

(1- Loudi Technician College, Loudi 417000;
2- College of Mathematical Sciences, Guangxi Normal University, Guilin 541004;
3- Department of Pharmacy, Baotou Medical College, Baotou 014040;
4- Lijiang College, Guangxi Normal University, Guilin 541004)

Abstract: The comparison of differences of populations is an important research topic in medical studies, economical and educational fields. This paper studies the construction of the quantile differences of response variables in two linear models with missing data. The fractional linear regression imputation method is used to impute the missing data of the response variables, and 'complete' data for two linear regression models are obtained. The empirical log-likelihood ratios of quantile differences of response variables are constructed based on the imputed data. Under some mild conditions, it is proved that the asymptotic distributions for the empirical log-likelihood ratios are scaled χ_1^2. The empirical likelihood confidence intervals for quantile differences of the response variables are then constructed based on this results. Simulations show that fractional imputation can improve the coverage accuracy of confidence intervals.

Keywords: linear model; quantile; fractional linear regression imputation; empirical likelihood; confidence intervals

Received: 14 Dec 2007. **Accepted:** 18 Sep 2008.

Foundation item: The National Natural Science Foundation of China (10971038); the Science Foundation of Guangxi (0728092); the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry ([2004]527); the Innovation Project of Guangxi Graduate Education ([2006]40).